dialectica

Vol 77 : 1
March 2023

The Minimalist Theory of Truth and the Generalisation Problem

Li Zhang & Leon Horsten

doi:10.48106/dial.v77.i1.02

Li Zhang & Leon Horsten. 2023. "The Minimalist Theory of Truth and the Generalisation Problem." *Dialectica* 77(1): 1–1. doi:10.48106/dial.v77.i1.02.

EUROPEAN
SOCIETY
FOR
ANALYTIC
PHILOSOPHY

The Minimalist Theory of Truth and the Generalisation Problem

Li Zhang & Leon Horsten

The Minimalist Theory of Truth must show how it can prove certain truth-involving generalisations. Horwich has proposed two solutions to this challenge over the past decades. The first of these invokes Hilbert's ω-rule, and is unacceptable. The second proposal can naturally be viewed in different ways. We show how this second proposal is naturally interpreted as a suggestion to solve the truth generalisation problem using uniform reflection rules. We also argue that this is indeed the right way for Horwich to respond to the truth generalisation problem.

Over the past three decades, Horwich's *minimalism* has been the most discussed deflationary truth theory. Generally speaking, this theory claims that everything about truth can be explained by the collection of underived and unproblematic instances of the equivalence schema.

$< ES >. < p >$ is true if p.

In the literature, the equivalence schema <ES> is also known as the Tarski-schema or T-schema; its instances are known as Tarski-biconditionals or T-sentences. The theory consisting of all underived, unproblematic Tarski-biconditionals, namely, the theory taking all such biconditionals to be *axioms*, is called the "*Minimalist Theory of Truth*" (MT).

Firstly, Horwich believes that truth is *non-substantial*, so we should not define truth with any *substantial* concept. Instead, the meaning of "is true" is given by the collection of underived, unproblematic instances of the T-schema. Horwich believes that "our understanding of"is true"—our knowledge of its meaning—consists in the fact that the explanatorily basic regularity in our use of it is the inclination to accept instantiations of the schema (E) "the proposition that *p* is true if and only if *p*" by declarative sentences of English (including any extensions of English)" (Horwich 1998, 35). Due to its nonsubstantiality, truth should remain neutral in debates in other philosophical and non-philosophical areas (Horwich 1998, 52).

Secondly, Horwich argues that MT alone suffices to explain all the truth-involving phenomena (Horwich 1998, 5). Thus, for instance, without equating truth with utility or any other substantial concept, MT suffices to explain that "true beliefs help us to achieve practical goals" (Horwich 1998, 44). In what follows, we denote the first point Horwich has made as *the neutrality thesis*, and the second as the *adequacy thesis* of minimalism (Gupta 1993, 361).

Despite Horwich's clever arguments for the two minimalistic theses, many logicians and philosophers insist that Horwich's minimalism is far from correct, since there are many truth-involving facts that cannot be explained by it. In particular, it cannot prove truth generalisations that we regard as acceptable. For instance, it is unclear how MT proves:

1. Every proposition of the form $p \rightarrow p$ is true.

Or

2. Every proposition is such that either it or its negation is true.

In fact, many believe that it is impossible for MT to prove sentences such as (1) and (2). In the literature, this problem is known as *the truth generalisation problem* (Halbach 2014, 57; Raatikainen 2005, 177). Horwich has formulated two proposals in response to this challenge, but, as they stand, neither of them decisively answers the problem. We will defend an amplification and extension of Horwich's second proposal, and argue that this successfully tackles the truth generalisation problem within the framework of truth-theoretic minimalism.

The structure of this paper is as follows: in section 1, we reformulate Horwich's minimalist truth theory in such a way that some unclarities of his original formulation are removed. In section 2, we show why MT and its modifications cannot prove intuitively acceptable truth generalisations. In section 3, we evaluate Horwich's two proposals in the light of critiques of them that have appeared in the literature. In section 4, we show how MT proves an ample collection of truth generalisations when strengthened with uniform reflection rules, and we argue this to be in line with Horwich's second proposal. In section 5, we conclude this paper by suggesting that Horwich should accept our formulation of the reflection rules proposal since it coheres best with his other truth-theoretic theses.

1 Reformulating MT

It has been recognized that several aspects of the formulation of MT are unclear. In particular, it is not clear which Tarski-biconditionals belong to MT's axioms, and which do not. Moreover, it is not clear how taking propositions as primary truth bearers increases MT's proof-theoretic strength. Thus, we suggest two modifications of MT in this paper. First, by applying the T-schema to sentences *that themselves do not contain the truth predicate*, we obtain a precise description of MT's axioms. Second, we take sentences to be primary truth bearers. Given these two modifications, MT is equivalent to the axiomatic truth theory TB (for "Tarski-biconditionals") when we take the Peano Arithmetic to be its base theory.[1]

One reason for MT's vagueness is Horwich's approach to truth-theoretic paradoxes. Horwich concedes that if some instances of the T-schema are included in MT's axioms, MT proves contradictions. He demonstrates this in the familiar way by applying the T-schema to the sentence:

> THE PROPOSITION FORMULATED IN CAPITAL LETTERS IS NOT TRUE.

He argues that the only acceptable strategy for this problem is to exclude some instances of the T-schema from the axioms of MT (Horwich 1998, 40–41). The spirit of his approach to paradoxes has been shared by prominent logicians, including Tarski: by putting different constraints on the scope of the T-schema, we obtain different formal truth theories. These theories capture central uses of the truth predicate, while in the meantime being adroit at avoiding truth-theoretic paradoxes. What renders Horwich's strategy different is that he does not give a specification of either the permitted or prohibited instances of the T-schema; he only requires that the collection of MT's axioms should be a maximally consistent set of sentences (Horwich 1998, 42). Unfortunately, McGee has shown there are uncountably many mutually incompatible sets that satisfy this requirement; none of them are recursively axiomatisable. Therefore, Horwich must impose more constraints on the instances of the T-schema (McGee 1992, 236–237).

TB is axiomatisable and consists of unproblematic Tarski-biconditionals, which renders it a suitable substitute for MT. However, far be it from us to

[1] TB is also sometimes denoted as DT (for "disquotational theory") in the literature (Halbach 2014, 53).

doi: 10.48106/dial.v77.i1.02

claim that TB is the *only* suitable substitute for MT. Many natural axiomatic disquotational theories of truth would do just as well. For instance, if one would substitute variants of Halbach's theory of *Positive Tarski-biconditionals* (Halbach 2001, 5) for MT instead, then the arguments of the present article would still go through.

Most logicians who are interested in formal truth theories, such as Tarski (Tarski 1944, 342), McGee (McGee 1992, 235), Halbach (Halbach 2014, 12) and Cieśliński (Cieśliński 2018, 1083, fn. 8), take sentences to be primary truth bearers. The reasons for their choice are quite straightforward: propositions are ill-understood and controversial, whereas we have rigorous and widely accepted syntactical theories of sentences.

Horwich nonetheless insists on formulating minimalism in terms of propositions because he believes that there exist propositions that cannot be expressed by current human languages (Horwich 1998, 20–21, fn. 4):

> Patrick Grim pointed out to me that the minimal theory cannot be regarded as *the set* of propositions of the form $<<p>$ is true if $p>$; for there is no such set. The argument for this conclusion is that if there were such a set, then there would be distinct propositions regarding *each* of its subsets, and then there would have to be distinct axioms of the theory corresponding to those propositions. Therefore there would be a 1-1 function correlating the subsets of MT with some of its members. But Cantor's diagonal argument shows that there can be no such function. Therefore, MT is not a set. In light of this *result* [our emphasis], when we say things like "$<A>$ follows from the minimal theory," we must take that to mean, not that the relation of *following from* holds between $<A>$ and a certain entity, the minimal theory; but rather that it holds between $<A>$ and *some part* of the minimal theory—i.e., between and some set of propositions of the form $< <p>$ is true if $p >$.

The particular argument of Grim that is alluded to here goes as follows.[2] Suppose there were a set S of all truths, and consider all subsets of S, i.e., all members of the power set $\mathcal{P}(S)$. To each element of this power set will correspond a truth. To each element of the power set, for example, a particular truth p either will or will not belong as a member. In either case, we will have

[2] See (Grim 1988).

a truth: that p is a member of that element, or that it is not. There will then be at least as many truths as there are elements of the power set $\mathcal{P}(S)$. But by Cantor's theorem, we know that the power set of any set will be larger than the original. There will then be more truths than there are members of S, and for any set of truths S there will be some truth left out. There can therefore be no set of all truths.

The quotation by Horwich shows that he regards Grim's argument as definitive: he takes the conclusion of the argument as a philosophical *result*. But it is far from clear whether, in the absence of a detailed, widely accepted theory of propositions and their constituents, Grim's argument is persuasive. To give but one example of a worry that one might have here,[3] observe that Grim's succinct argument presupposes that for each subset B of S, there exists a *proposition* of the form $p \in S$ ($\neg p \in S$). For this to be the case, for each such subset B there has to be an individual concept of B as a part of this proposition. But whether all such individual concepts exist, is a substantial and unsettled philosophical question.

In view of this, there seems to be no pressing need for Horwich to take truth to be a property of propositions. Nonetheless, we do not ask of Horwich that he abandons his views of the kinds of entities that are the bearers of truth. A sub-class of the totality of all propositions is the *set* of all propositions that can be expressed by sentences belonging to some fixed language. The theory of true *sentences* (of some language) can then be seen as a special case of Horwich's more general theory of truth of propositions. So the argument that is developed in the subsequent sections intends to support the thesis that, *as far as truth of (propositions expressed by) sentences goes*, Horwich's arguments of the early 2000s concerning truth generalisations were at least a decade ahead of their time, albeit not fully fleshed out. That Horwich might well be sympathetic to such an interpretation of his views concerning truth generalisations is indicated by the passage in his *Truth* book, where he says that:

> […] ordinary language suggests that truth is a property of propositions, and that utterances, beliefs, assertions, etc., inherit their truth-like character from their relationship to propositions. However, [previous considerations] show that this way of seeing things has no particular explanatory merit. The truth-like conception for

[3] We do not have the space to go deeply into the literature that has been generated by Grim's argument.

doi: 10.48106/dial.v77.i1.02

> each type of entity is equally minimalistic. And by assuming any one of them we can easily derive the others. (Horwich 1998, 102)

2 The Truth Generalisation Problem

A non-trivial general claim of the form "*every* x is φ" cannot be proved by a finite collection of premises each of which asserts that a_i is φ, for some i, except if there is an additional premise that says that every object is one of this finite number of a_i's. This also applies to MT in the sense that a general claim of the form "for *every* sentence x of the form $p \to p$, x is true," for example, cannot be proved in MT. Indeed, it has been *proved* that such a truth generalisation cannot be proved in TB (=MT) (Halbach 2014, 56–57). Many such truth generalisations appear to be conceptual truths about the concept of truth. In particular, this is so for the classical compositional axioms of truth that state that truth commutes with the logical connectives. Moreover, there are valid philosophical and natural language arguments whose *validity* depends not just on Tarski-biconditionals, but also on compositional truth axioms (Fischer 2023, CHANGE PAGENUMBER, WAS 1; Hilbert 1931, 5).

This poses a challenge to the Minimalist Theory of Truth: recall that MT's adequacy thesis claims that *all* facts whose expression involves the truth predicate can be explained by assuming no more about truth than instances of the equivalence schema (Horwich 1998, 23). A number of philosophers and logicians, including Armour-Garb (Armour-Garb 2010, 698), Gupta(Gupta 1993, 363–364), Halbach (Halbach 2001, 1959–1960) and Soames (Soames 1997, 30–31) regard its inability to prove truth generalisations as a serious defect of MT.

One may be tempted to appeal to "McGee's trick" (McGee 1992, 238), and contend that since it is always possible to find a T-sentence that is equivalent to a given truth generalisation, when MT is not identified with TB but instead with TB plus such additional Tarski-biconditionals, MT is capable of proving all acceptable truth generalisations. Indeed, by the diagonal lemma, for every truth generalisation A, there is a sentence $\varkappa$ such that:

$$\vdash \varkappa \leftrightarrow (T(\varkappa) \leftrightarrow A).$$

By associativity of ↔, the Tarski-equivalence $\varkappa \leftrightarrow T(\varkappa)$ is provably equivalent to A, where A is an acceptable truth generalisation. However, it is widely accepted that sentences such as that expressed by $\varkappa$ should not be allowed in , since by exactly the same procedure, it is possible to find a T-sentence equivalent to "Santa Claus exists," which should not follow from any acceptable truth theory.

In sum, the generalisation problem poses a serious challenge for Horwich's truth theory.

3 Horwich's Responses

It is not clear exactly when Horwich came to the conclusion that MT cannot prove acceptable truth generalisations. But it is clear he wants to resolve this problem by strengthening MT with further theoretical resources. Moreover, it is possible to group his many responses into two categories: the *ω-rule proposal* and a *reflection-based proposal*. We review these proposals in turn.

3.1 *Horwich's First Attempt: the ω-rule*

In the postscript of the revised *Truth*, Horwich formulates his first attempt at solving the truth generalisation problem. There he writes:

> However, it seems to me that in the present case, where the topic is *propositions*, we can find a solution to this problem. For it is plausible to suppose that there is a truth-preserving rule of inference that will take us from a set of premises attributing to each proposition some property, *F*, to the conclusion that all propositions have *F*. No doubt this rule is not *logically* valid, for its reliability hinges not merely on the meanings of the logical constants, but also on the nature of propositions. But it is a principle we do find plausible. We commit ourselves to it, implicitly, in moving from the disposition to accept any proposition of the form "*x* is *F*" (where *x* is a proposition) to the conclusion "All propositions are *F*." So we can suppose that this rule is what sustains the explanations of the generalizations about truth with which we are concerned. Thus we can, after all, defend the thesis that the basic theory of truth consists in some subset of the instances of the equivalence schema. (Horwich 1998, 137–138)

It has been acknowledged that the above mentioned truth-preserving rule amounts to a form of the ω-rule (Raatikainen 2005, 175). Hilbert introduces this principle in the following manner:

> If it has been proved, for any given numeral δ, that the formula
>
> $$\mathfrak{A}(\delta)$$
>
> is always a correct numerical formula, then the formula center
>
> $$(x)\mathfrak{A}(x)$$
>
> can be laid down as a starting formula [Ausgangsformel]. (Hilbert 1931, 1154)

Feferman rightly observed that Hilbert's own formulation of the ω-rule is somewhat vague (Feferman 1986, 212). The ω-rule is perhaps more clearly expressed as: "From infinitely many premises $\varphi(0)$, $\varphi(1)$, … that result from replacing the numerical variable n in $\varphi(n)$ with the numeral for each natural number, conclude $\forall x\varphi(x)$" (Hazen 1998).

The ω-rule is a strong rule: When enriched with this rule, PA proves true arithmetic (Hazen 1998). With regard to the generalisation problem, when augmented with the ω-rule, MT is able to prove all acceptable truth generalisations. Take a finite first-order language as an example; every sentence of the form $p \rightarrow p$ is a theorem of this language. Enumerate all sentences of the form $p \rightarrow p$, so each of them is represented by a numeral. Apply T-sentences of MT to them, so for each n, $T(n)$. By the ω-rule, we obtain the general claim $\forall xT(x)$.

However, certain features of the ω-rule render this proposal problematic, and in particular unacceptable to the minimalist truth theory. Raatikainen has argued that we, as finite human beings, cannot take infinitely many premises into consideration simultaneously. Therefore, even if the theory MT + the ω-rule is capable of proving acceptable truth generalisations, those generalisations are beyond the reach of ordinary human beings (Raatikainen 2005, 176). This problem with the ω-rule cannot be overcome: it simply has no effective (read: recursively enumerable) equivalent. Moreover, the proof-theoretic strength of the ω-rule makes it specifically unacceptable to the minimalist truth theory. When enriched with this rule, Peano Arithmetic proves all true arithmetic sentences. True arithmetic is not axiomatisable, while MT is intended to be an axiomatised truth theory.

It is not clear whether or not Horwich has accepted critiques of his first proposal. In a recent publication Horwich still seems to propose using the ω-rule as a solution to the truth generalisation problem:

> For it is plausible to suppose that there is a truth-preserving rule of inference that will take us from a set of premises attributing to each proposition of a certain form some property, G, to the conclusion that the *all* proposition have property G. And this rule – not *logically* valid, but nonetheless necessarily truth-preserving given the nature of proposition – enables the general facts about truth to be explained by their instances. (Horwich 2003, 84, fn. 14)

Yet in most of his recent writings, Horwich advocates an alternative resolution, based on an introspective process. To this proposal we now turn.

3.2 *Horwich's Second Attempt: reflection*

Over the years, Horwich's formulation of his second proposal has varied, and it is not easy to select a preferred formulation from these variants. Extant critiques of his various formulations are indecisive. Nonetheless, we will argue that all variants of Horwich's second proposal need emendation in order to solve the truth generalisation problem.

A first fomulation of Horwich's second attempt emerges in (Horwich 2001), which appeared in 2001:

> Whenever someone can establish, for any *F*, that it is *G*, *and recognizes that he can do this*, then he will conclude that every *F* is *G*. (Horwich 2001, 157)

Call this *Solution 2.0*. This solution also consists in adding an additional rule of inference to MT. But the additional rule of inference of Solution 2.0 is different from the ω-rule.

In a revised version (2010) of the same paper, Horwich formulates a variant of this new proposal, which in effect amounts to a further, *substantially different* proposal:

> Whenever someone is disposed to accept, for any proposition of structural type F, that it is G (and to do so for uniform reasons) then he will be disposed to accept that every F-proposition is G (Horwich 2010, 45).

To the above statement, he adds the following *proviso*:

doi: 10.48106/dial.v77.i1.02

> We cannot conceive of there being additional Fs – beyond those Fs we are disposed to believe are G – which we would not have the same sort of reason to believe are Gs (Horwich 2010, 44–45).

Call the proposal that is encapsulated in the previous two quotations *Solution 2.1*. (Horwich endorses this same solution in 2005 (Horwich 2003, 84).)

Armour-Garb argues that Solution 2.1 is unsatisfactory because:

> One will not be disposed to accept (the proposition) that all F-propositions are G, from the fact that, for any F-proposition, she is disposed to accept that it is G (NB, even for uniform reasons), unless she is *aware* of the fact that, for any F-proposition, she is disposed to accept that it is G. (Armour-Garb 2010, 699)

The *proviso* that Horwich added to Solution 2.1 does not provide such an awareness component. It merely adds a *negative* condition ("not being able to conceive of there being F's that are not G"), while Armour-Garb's awareness-requirement is a positive condition. Nonetheless, Solution 2.0 incorporates exactly the awareness condition that Armour-Garb insists on ("and recognises that he can do this").

Armour-Garb is making a psychological observation here, but there is an accompanying *rational* point to be made also. If one does not *recognise* that for any F-proposition, she is disposed to accept that it is G, then she is not, without further ado, *rationally required* to believe that every F-proposition is G. *Ought* implies *can*, and in this situation she simply lacks the ground for accepting that every F-proposition is G.[4] For this reason, Horwich's Solution 2.0 must be regarded as superior to his Solution 2.1.

Nonetheless, Armour-Garb would not be satisfied with Solution 2.0 either. He argues that the switch, in the move from the premise to the conclusion of the rule of inference in Solution 2.1, of "for any F-proposition" from outside the "disposed to accept"-context to inside the "disposed to accept"-context, is "viciously circular." He is certainly right that this quantifier shift, which is also present in Solution 2.0, is not derivable in classical logic. Nonetheless, we take issue with this aspect of Armour-Garb's critique of Horwich's second proposal. Indeed, we agree with Cieśliński that Armour-Garb's dismissal of Horwich's second solution on the ground of its being viciously circular is "hasty" (Cieśliński 2018, 1082): we will come back to this later.

[4] Further discussion of these important matters can be found in [UNKNOWN REFERENCE].

It is time to spell out the content of Horwich's Solution 2.0, i.e., the first quotation in this section, in more precise terms. We do this by formalising Horwich's informally expressed—and somewhat vague–ω-rule in first-order logic. In our formalisation of the first quotation in this section, we want to be charitable to Horwich. We do not claim that Horwich would agree with our formalisation (Horwich can speak for himself), but we will argue that there are good reasons for him to do so. Firstly, Solution 2.0 contains the phrase "*will* conclude," making it seem like a psychological prediction.[5] If it is taken in this way, then whether it is true or not is an empirical matter. But this is presumably not what Horwich intends. Rather, what he means, is that the agent will be disposed to draw this conclusion *if she is rational*. In other words, Horwich purports to propose a rational *rule of inference* here. So it might be better to replace, in Solution 2.0, "will conclude" by "may (rationally) conclude," or perhaps even "should (rationally) conclude." Secondly, since we are concerned with *establishing* truth generalisations, we identify the concepts "being disposed to accept" and "recognising" with being *provable*. In particular, we interpret the clause "*and recognizes that he can do this*" as *de re* provability of an arbitrary F *that* it is G. Thirdly, we identify provability with provability in the background theory, which is MT. If we were to identify provability with provability in the system *including the rule*, then the proposed rule would indeed be viciously circular, confirming Armour-Garb's (unfounded) suspicions. But if we identify provability with provability in MT, then there is no circularity. Fourthly, we *omit* the concept of provability ("being disposed to accept") from the conclusion of the rule. With these precisifications in place—which we take to be reasonable, but we leave it open whether they are *exactly* in accordance with what Horwich intended—we obtain the following schematic rule:[6]

$$\frac{\vdash \forall x : F(x) \rightarrow Bew_{MT}(G(x))}{\vdash \forall x : F(x) \rightarrow G(x)}.$$

We will call this rule H (for: "Horwich"). Observe that, unlike the ω-rule, H is an *effective* rule: adding it to MT yields an axiomatic system.

[5] Cieśliński sees this as the main weakness of Horwich's recent views: see (Cieśliński 2017, 80).

[6] In the interest of readability, we are sloppy with Gödel coding here as well as later on in this article.

Worries based on the lottery paradox might cause one to doubt the rationality of rule H. For any ticket (in a large, fair lottery), I believe that it is not the winning ticket (and I believe this for "uniform reasons"). But from this, I am not prepared to infer that every ticket is a losing ticket (Kyburg 1970, 56). Nonetheless, such a worry would be ill-founded, for the situation under consideration is different in one key respect. The irrationality of the lottery paradox inference stems from the fact that many small but non-zero probabilities (of being the winning ticket) can add up to a large probability (of one of a large collection of tickets being the winning one). But what is provable has probability 1 rather than $1 - \epsilon$ (for some small ϵ), since provability in a sound system from necessary premises is itself necessary, and necessary truths by a Kolmogorov axiom for probability receive probability 1. So the fair lottery phenomenon is irrelevant to the evaluation of rule H.[7]

4 Uniform Reflection and Truth Generalisations

We have seen that Horwich recognises that H is not an admissible inference rule of first-order logic. The main questions that we want to answer in this section about H are the following: *To what extent and in which way does adding H to MT allow us to prove truth generalisations?* Moreover: *Is H a rational rule of inference?*

4.1 *H and Uniform Reflection*

It is clear that given a *sound* theory S, adding H (with Bew_{MT} replaced by Bew_S) to S, results in a sound system. So, in particular, MT + H is a sound system.

Next, we make the crucial observation that H is equivalent to a reflection rule that has been intensively investigated in proof theory. To this end, we first recall the notion of *uniform reflection principle* for a theory S (denoted as $RFN(S)$),

$$\forall x : Bew_S(\varphi(x)) \to \varphi(x),$$

and the notion of *uniform reflection rule* for a theory S (denoted as UR_S),

[7] An extended discussion of the relevance or irrelevance of the lottery paradox in this context can be found in (Cieśliński 2017, sec. 13.5).

$$\frac{\vdash \forall x : Bew_S(\varphi(x))}{\vdash \forall x : \varphi(x)}.$$

Feferman has proved the remarkable little fact that RFN(S) is equivalent to UR_S (Feferman 1962 Theorem 2.19). In the light of this, it is easy to see that H is equivalent to UR_{MT} (and therefore also to RFN(MT)): the ⇒-direction is obvious, and the ⇐-direction follows immediately from Feferman's theorem.

At this point, a connection with Horwich's *first* solution also becomes apparent. Indeed, the uniform reflection rule is widely seen as an effective version (a "tamed" version) of the ω-rule. Horwich's appeal to the ω-rule was (rightly) rejected by Raatikainen on account of its non-effectiveness. Uniform reflection rules cannot be rejected on the same grounds.

We will now see how the main observation of this subsection allows us to answer the question to what extent H enables us to prove truth generalisations.

4.2 *Deriving truth generalisations*

Let us denote MT + H as MT_1. Now that we have made Horwich's Solution 2.0 precise, we address the question whether MT_1 can prove all intuitively acceptable truth generalisations. An apparent counterexample is a proposition such as "there are as many truths as there are untruths" (Gupta 1993, 363). But this is a second-order statement, involving not just sentences but also *sets* of sentences. So it falls outside the scope of MT (=TB), which cannot even express *claims* involving sets of sentences.

The truth theory that takes the axioms that state that truth commutes with the logical connectives for sentences that do not themselves contain the notion of truth, is called CT. It is fairly generally accepted that in CT, a vast amount of intuitively acceptable truth generalisations logically follow [UNKNOWN REF, chapter 6]. So if Horwich can derive the truth axioms of CT, then he has made significant progress towards solving the truth generalisation problem. Nonetheless, it would be an exaggeration to say that *all* intuitively acceptable truth generalisations are provable in CT:[8] the truth generalisation "All arithmetical theorems of CT are true," for instance, is not provable even in CT.

[8] Thanks to an anonymous referee for making this point.

doi: 10.48106/dial.v77.i1.02

With only one exception, the compositional truth axioms of CT can indeed be derived in MT_1 (Horsten and Leigh 2017). As an example, let us consider the compositional axiom for negation:

$$\forall x \in \mathcal{L}_{PA} : T(\neg x) \leftrightarrow \neg Tx.$$

Every *instance* of this axiom can be proved in TB (using Tarski-biconditionals). Moreover, *that* every instance can be proved in TB, can be uniformly recognised (i.e., proved) as a combinatorial fact even in the background theory PA. So we have:

$$PA \vdash \forall x \in \mathcal{L}_{PA} : Bew_{MT}(T(\neg x) \leftrightarrow \neg Tx).$$

Then by UR_{MT} we indeed obtain $\forall x \in \mathcal{L}_{PA} : T(\neg x) \leftrightarrow \neg Tx$.

The other compositional axioms can be derived in a similar way in MT_1, with the sole exception of the quantifier axiom:

$$\forall \varphi(x) \in \mathcal{L}_{PA} : T(\forall x \varphi(x)) \leftrightarrow \forall x T\varphi(x).$$

We cannot prove in MT, for every $\varphi(x) \in \mathcal{L}_{PA}$, that $T(\forall x\varphi(x)) \leftrightarrow \forall x T\varphi(x)$. The reason is that TB (=MT) only contains Tarski-biconditionals for *sentences*, i.e., for *closed* formulas. In order to prove, for each $\varphi(x) \in \mathcal{L}_{PA}$, that $T(\forall x\varphi(x)) \leftrightarrow \forall x T\varphi(x)$, we need a slight strengthening of the Tarski-biconditionals of TB, namely the *uniform* arithmetical Tarski-biconditionals, which are the sentences of the form $\forall x(T\varphi(x) \leftrightarrow \varphi(x))$, for formulas $\varphi(x) \in \mathcal{L}_{PA}$. The resulting slight strengthening of TB is called UTB.

How do we derive these uniform Tarski-biconditionals? We can prove them in MT_1 as follows [Horsten and Leigh (2017), Theorem 9][9]. Every instance of a given uniform (arithmetical) Tarski-biconditional can be proved in TB. This combinatorial fact can again be proved even in PA :

$$PA \vdash \forall x \in \mathcal{L}_{PA} : Bew_{MT} : T\varphi(x) \leftrightarrow \ \varphi(x).$$

So by applying UR_{MT} in MT_1 to this fact, we obtain the result. Now, in a second stage, we can proceed as we did with the negation axiom. But to carry out this proof, we need to appeal to UR_{MT_1}, which is the uniform reflection rule for MT_1:

[9] Theorem 9 obtained in (Horsten and Leigh 2017) is based on uniform reflection principles rather than rules, but we have seen above that by an argument due to Feferman, the two are provably equivalent.

$$\frac{\vdash \forall x : Bew_{MT_1}(\varphi(x))}{\vdash \forall x : \varphi(x)},$$

Where φ can be any arithmetical formula, and Bew_{MT_1} formally expresses provability in MT_1. For the same reasons as why UR_{MT} exceeds MT, the rule UR_{MT_1} exceeds MT_1. If we apply this inference rule to the earlier obtained fact that PA proves:

$$\forall x \in \mathcal{L}_{PA} : Bew_{MT_1}(T(\forall x\varphi(x)) \leftrightarrow \forall xT\varphi(x)),$$

Then we obtain the desired result that $\forall x \in \mathcal{L}_{PA} : T(\forall x\varphi(x)) \leftrightarrow \forall xT\varphi(x)$.

In sum, we can prove all the compositional truth axioms of CT, and therefore many intuitively acceptable truth generalisations in $MT_2 = MT_1 + UR_{MT_1} = MT + UR_{MT} + UR_{MT+UR_{MT}}$ (Horsten and Leigh 2017). In other words, many truth generalisations follow by two iterations of uniform reflection on MT. Even more truth generalisations can be proved when this strategy is iterated further. By adding further uniform reflection principles to MT_2, for instance, also the truth generalisation "All arithmetical theorems of CT are true" become provable.

At this point, we see that we have to go slightly beyond our charitable interpretation of Horwich's Solution 2.0. Horwich claims that *one* level of reflection on MT suffices to prove all acceptable truth generalisations. We now see that *two* levels of reflection on MT are required. Given the equivalence between Horwich's rule H and Feferman's uniform reflection rule, all acceptable truth generalisations can be derived in the theory MT+H+H', where H' is just like H, except that its background theory is MT+H instead of MT:

$$\frac{\vdash \forall x : F(x) \rightarrow Bew_{MT+H}(G(x))}{\vdash \forall x : F(x) \rightarrow G(x)}.$$

In sum, if H and H' are *rational* rules of inference, then Horwich was very much on the right track.

4.3 *Rationality*

Uniform reflection rules are rules that contain the required "awareness" component in the antecedent (the agent has to have a proof) and that are also, *pace* Armour-Garb, not circular in any way. In addition, in the premise of uniform reflection rules, the awareness/recognition component that is required is *proof* from the Tarski-biconditionals.

On our interpretation and emendation of his view, Horwich contends that it is *rational* to add UR_{MT} and UR_{MT_1} to MT. With this, he would not be alone. In his work on *implicit commitment*, Feferman claimed that if an agent explicitly accepts a theory S, then she also ought to accept uniform reflection principles and rules for S, such as UR_S and UR_{S+UR_S} (Feferman 1991, 2, 44). Acceptance of UR_S, is, in his view, *implicit* in acceptance of S, and acceptance of UR_{S+UR_S} is "implicitly implicit" in the acceptance of S.

Feferman did not give an epistemological argument for *why*, if one accepts a theory S, one should also accept UR_S (and UR_{S+UR_S}). A recent attempt to provide such an argument is given by Fischer in (Fischer 2023), which can, in retrospect, be seen as one attempt to develop Horwich's Solution 2.0 in detail. A discussion of Fischer's argument is outside the scope of this article. Here, we restrict ourselves to a few remarks on the issue. The uniform reflection rule for the theory that one is currently working in expresses a form of trust or confidence in this theory. If the theory one is working in is justified, then this trust is also justified, and therefore accepting the uniform reflection rule is justified. The theory that is relevant in the present context is the truth theory MT. Horwich argues that this theory is indeed justified, because Tarski-biconditionals express the content or meaning of the concept of truth (Horwich 2010, 17). Therefore, making one's trust in MT explicit by accepting UR_{MT} and UR_{MT_1} is rational.[10] Since, by Feferman's theorem, H is equivalent to UR_{MT}, and H' is equivalent to UR_{MT_1}, H and H' are therefore also rational inference rules.

5 Horwich Vindicated?

There have been two phases in the history of truth-theoretic deflationism. In the first phase, disquotational axioms were taken to express the full content of the concept of truth. This phase comprises, a.o., Quine's views on truth as a tool for semantic ascent and descent (Quine 1970, 10–13), and the prosentential theory of truth (Grover, Camp and Belnap 1975). Horwich's minimalism is often viewed as a late and particularly bright exponent of this phase of deflationism. In the second phase, compositional axioms were taken to ex-

[10] Considerations such as these may provide at least the beginnings of a response to Cieśliński's complaint above (cfr supra) that Horwich's theory is too psychological.

press basic properties of the concept of truth. This phase started sometime in the 1980s, partly under the influence of Davidson's truth-conditional compositional approach to natural language semantics (Davidson 1967). During much of this second phase, Horwich's views on the concept of truth came to be increasingly seen as dated and untenable. As a result of this, his writings about the generalisation problem after the first edition of his book *Truth* were mostly ignored by the truth-theoretic community.

Perhaps we now experience the dawn of a third phase in the history of truth-theoretic deflationism, in which the relation between the concept of truth on the one hand, and reflection principles on the other hand, play a major role. In particular, it is currently a hotly debated question whether, by making use of reflection principles or rules, disquotationalism can solve the generalisation problem. We make no attempt to adjudicate this discussion here. But we have seen that Horwich anticipated the current philosophical debate already in the early 2000s. So rather than being a truth-theoretic dinosaur, at the time Horwich's views were ahead of their time—which of course does not mean that they are in any way definitive.

The main reason why Horwich's thoughts about the relation between reflection principles and truth generalisations were ignored is that Horwich's view about this problem was not completely precise and was connected to other views of his that can be separated from the problem at issue. Horwich was committed to propositions as the bearers of truth, but did not give a precise theory of propositions. At the same time, he was also committed to the background disquotational theory as a maximal consistent collection of propositions, which prevents it from being recursively axiomatisable, and therefore prevents it from being learnable. But we have seen that a *derived* notion of true proposition *expressible in a given language* makes perfect sense in Horwich's framework. Moreover, Horwich's requirement of MT being a maximal consistent collection of propositions is unrelated to his solution proposal to the generalisation problem, and can therefore simply be rejected—which is exactly what the truth-theoretic community has largely done. In sum, Horwich's views from the early 2000s on the truth generalisation problem can be disentangled from the further commitments and unclarities with which he connected them.

The imprecision of his treatment of the generalisation problem prevented Horwich from working out the technical details with full precision. For instance, he did not see that *two* rounds of uniform reflection are needed in order to derive the compositional truth principles from the disquotational axioms.

Nonetheless, Horwich did see that his strategy for dealing with the generalisation problem is in line with his two main minimalistic theses: the neutrality thesis and the adequacy thesis. Reflection rules are not truth-theoretic (or: *philosophical*), but *mathematical* rules. Uniform reflection rules are universally seen as mathematical rules because they have substantial mathematical consequences; they are canonical ways for extending the mathematical strength of a theory. Therefore, strengthening MT with uniform reflection rules does not affect the neutrality of the theory of *truth*. (Indeed, as mentioned earlier, MT can be taken to be proof-theoretically conservative over its background theory PA.) Moreover, since CT is derivable from MT by means of two rounds of uniform reflection, and CT proves the needed truth generalisations, a solution to the generalisation problem is reached, whereby the challenge to the adequacy thesis is answered.

The more recent debate about the connection between reflection principles and the truth generalisation problem developed only after 2015, and it developed largely independently from Horwich's views on the generalisation problem. Moreover, we now see further and more clearly in these matters than Horwich did around 2002. Yet it would be a mistake to take Horwich's early thoughts on this issue to be merely of historical relevance ("give credit where credit is due"). The appeal to proof theoretic reflection principles and rules as a means to derive compositional truth axioms is sometimes seen as a mere "technical" manoeuvre. But Horwich, at the time, did not know any of the proof theoretic literature concerning reflection principles and hit on the basic idea *in tempore non suspecto*. Purely by philosophically thinking about how to solve the generalisation problem in a disquotational framework, he, in one of his proposals (Proposal 2.0), arrived at the view that the compositionality or truth follows by the uniform reflection rule from disquotational principles. This is simply amazing, and it shows that rather than being merely a technical trick, it is a very natural theoretical view to take.[*]

Li Zhang
0000-0002-7766-7263

[*] The authors are indebted to the members of the Antos-Horsten doctoral seminar at the University of Konstanz, where an early version of this article was presented, for helpful and thoughtful comments. They also thank two anonymous referees, whose comments have led to substantial improvements of this article. The first author's research for this article was supported by a Chinese Scholarship Council Fellowship (CSC No. 201906210194) for carrying out doctoral research at the University of Bristol and at the University of Konstanz. Without the support of this scholarship, the present article could not have been written.

Tsinghua University
l-zhang17@mails.tsinghua.edu.cn

Leon Horsten
0000-0003-3610-9318
Universität Konstanz
Leon.Horsten@uni-konstanz.de

References

ARMOUR-GARB, Bradley. 2010. “Horwichian Minimalism and the Generalization Problem.” *Analysis* 70(4): 693–703, doi:10.1093/analys/anq073.

BEALL, J. C. and ARMOUR-GARB, Bradley, eds. 2004. *Deflationism and Paradox*. Oxford: Oxford University Press, doi:10.1093/oso/9780199287116.001.0001.

CIEŚLIŃSKI, Cezary. 2017. *The Epistemic Lightness of Truth. Deflationism and its Logic*. Cambridge: Cambridge University Press, doi:10.1017/9781108178600.

—. 2018. “Minimalism and the Generalisation Problem: on Horwich’s Second Solution .” *Synthese* 195(3): 1077–1101, doi:10.1007/s11229-016-1227-5.

DAVIDSON, Donald. 1967. “Truth and Meaning.” *Synthese* 17(1): 304–323. Reprinted in Davidson (1984, 17–36), doi:10.1007/BF00485035.

—. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, doi:10.1093/0199246297.001.0001.

EWALD, William Bragg, ed. 1996. *From Kant to Hilbert: A Source Book in the Foundations of Mathematics. Volume II*. Oxford: Oxford University Press.

FEFERMAN, Solomon. 1962. “Transfinite Recursive Progressions of Axiomatic Theories.” *The Journal of Symbolic Logic* 27(3): 259–316, doi:10.2307/2964649.

—. 1986. “Introductory Note to 1931c [Gödel (1931)].” in *Collected Works. Volume I: Publications 1929–1936*, pp. 208–212. Oxford: Oxford University Press. Edited by Solomon Feferman, John W. Dawson, Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay and Jean van Heijenoort, doi:10.1093/oso/9780195072556.003.0005.

—. 1991. “Reflecting on Incompleteness.” *The Journal of Symbolic Logic* 56(1): 1–49, doi:10.2307/2274902.

FISCHER, Martin. 2023. “Another Look at Reflection.” *Erkenntnis* 88(2): 479–509, doi:10.1007/s10670-020-00363-9.

GÖDEL, Kurt. 1931. “Besprechung von Hilbert (1931).” *Zentralblatt für Mathematik und ihre Grenzgebiete* 1: 260. English translation in van Heijenoort (1967, 596–616); reprinted in Gödel (1986, 213/214).

—. 1986. *Collected Works. Volume I: Publications 1929–1936*. Oxford: Oxford University Press. Edited by Solomon Feferman, John W. Dawson, Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay and Jean van Heijenoort.

doi: 10.48106/dial.v77.i1.02

Grim, Patrick. 1988. “Logic and Limits of Knowledge and Truth.” *Noûs* 22(3): 341–367, doi:10.2307/2215708.

Grover, Dorothy L. 1992. *A Prosentential Theory of Truth*. Princeton, New Jersey: Princeton University Press, doi:10.1515/9781400862689.

Grover, Dorothy L., Camp, Joseph L., Jr. and Belnap, Nuel D., Jr. 1975. “A Prosentential Theory of Truth.” *Philosophical Studies* 27(1): 73–124. Reprinted as Grover (1992, ch. 3), doi:10.1007/bf01209340.

Gupta, Anil. 1993. “Minimalism.” in *Philosophical Perspectives 7: Language and Logic*, edited by James E. Tomberlin, pp. 359–369. Oxford: Basil Blackwell Publishers, doi:10.2307/2214129.

Halbach, Volker. 2001. “Disquotational Truth and Analyticity.” *The Journal of Symbolic Logic* 66(4): 1959–1973, doi:10.2307/2694987.

—. 2011. *Axiomatic Theories of Truth*. 1st ed. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511921049.

—. 2014. *Axiomatic Theories of Truth*. 2nd ed. Cambridge: Cambridge University Press. First edition: Halbach (2011), doi:10.1017/CBO9781139696586.

Hazen, Allen Patterson. 1998. “Non-Constructive Rules of Inference .” in *The Routledge Encyclopedia of Philosophy*, edited by Edward J. Craig. London: Routledge. The Routledge Encyclopedia was made available online in 2002 and is now regularly updated., doi:10.4324/9780415249126-Y014-1.

Hilbert, David. 1931. “Die Grundlegung der elementaren Zahlenlehre.” *Mathematische Annalen* 104(1): 485–494. Translated as “The Grounding of Elementary Number Theory” in Ewald (1996, 1148–1156), doi:10.1007/BF01457953.

Horsten, Leon and Leigh, Graham E. 2017. “Truth is Simple.” *Mind* 126(501): 195–232, doi:10.1093/mind/fzv184.

Horwich, Paul. 1990. *Truth*. Oxford: Basil Blackwell Publishers. Second edition: Horwich (1998).

—. 1998. *Truth*. 2nd ed. Oxford: Basil Blackwell Publishers. First edition: Horwich (1990), doi:10.1093/0198752237.001.0001.

—. 2001. “A Defense of Minimalism.” *Synthese* 126(1–2): 149–165. Substantially revised version in Horwich (2010, 35–56), doi:10.1023/a:1005279406402.

—. 2003. “A Minimalist Critique of Tarski on Truth.” in *Philosophy and Logic. In Search of the Polish Tradition. Essays in Honour of Jan Woleński on the Occasion of his 60th Birthday*, edited by Jaakko Hintikka, Tadeusz Czarnecki, Katarzyna Kijania-Placek, Tomasz Placek, and Artur Rojszczak, pp. 3–12. Synthese Library n. 323. Dordrecht: Kluwer Academic Publishers. Reprinted in Beall and Armour-Garb (2004, 75–84) and, in substantially revised form, in Horwich (2010, 79–97), doi:10.1007/978-94-017-0249-2_1.

—. 2010. *Truth – Meaning – Reality*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199268900.001.0001.

KYBURG, Henry E., Jr. 1970. "Conjunctivitis." in *Induction, Acceptance, and Rational Belief*, edited by Marshall SWAIN, pp. 55–82. Synthese Library n. 26. Dordrecht: D. Reidel Publishing Co., doi:10.1007/978-94-010-3390-9_4.

MCGEE, Vann. 1992. "Maximal Consistent Sets of Instances of Tarski's Schema (T)." *The Journal of Philosophical Logic* 21(3): 235–241, doi:10.1007/bf00260929.

QUINE, Willard van Orman. 1970. *Philosophy of Logic*. Cambridge: Cambridge University Press. Second edition: Quine (1986).

—. 1986. *Philosophy of Logic*. 2nd ed. Cambridge, Massachusetts: Harvard University Press. First edition: Quine (1970).

RAATIKAINEN, Panu. 2005. "On Horwich's Way Out." *Analysis* 65(3): 175–177, doi:10.1111/j.1467-8284.2005.00546.x.

SOAMES, Scott. 1997. "The Truth about Deflationism." in *Philosophical Issues 8: Truth*, edited by Enrique VILLANUEVA, pp. 1–44. Atascadero, California: Ridgeview Publishing Co., doi:10.2307/1522992.

TARSKI, Alfred. 1944. "The Semantic Conception of Truth and the Foundations of Semantics." *Philosophy and Phenomenological Research* 4(3): 341–375, doi:10.2307/2102968.

VAN HEIJENOORT, Jan, ed. 1967. *From Frege to Gödel: A Source Book in Mathematical Logic 1879-1931*. Cambridge, Massachusetts: Harvard University Press.